Anexo Técnico: Metodología, Validación y Caso de Negocio

Este documento constituye el apéndice técnico del informe final. Detalla la arquitectura de datos, la validación estadística de hipótesis, la formulación matemática de los modelos y la proyección financiera del proyecto SAREP.

A. Estrategia de Datos y Preprocesamiento

Para abordar el problema de **"Arranque en Frío"** (*Cold Start*) derivado de la falta de históricos centralizados y estructurados en la UNRC, se implementó una estrategia híbrida de adquisición de conocimiento:

1. **Análisis Descriptivo (Contexto):** Se procesaron los datos longitudinales de los informes de *Numeralia UNRC (2021-2025)* para diagnosticar la brecha de eficiencia terminal y establecer la línea base del problema (Sección 2.2.1 del informe principal). Se proceso cada categoría (Modalidad, Profesores, Desertores) en csv's y procesados con Pandas y Matplotlib la visualización básica descriptiva de la serie temporal trimestre por trimestre.
2. **Modelado Predictivo (Solución):** Se aplicó una técnica de **Adaptación de Dominio** (*Domain Adaptation*) utilizando un *Dataset Proxy Estandarizado* (UCI Machine Learning Repository) para entrenar el motor de inferencia inicial, validando posteriormente las características (*features*) más relevantes con datos locales.

A.1 Metodología del Análisis Longitudinal (Diagnóstico)

Ante la ausencia de históricos centralizados, se realizó una **reconstrucción forense de datos** longitudinales de la pagina oficial de (Numeralia,2021-2025) para establecer una línea base de desempeño institucional.

1. Fuentes y Recuperación de Datos: Se consolidaron registros dispersos en archivos estandarizados (`data/raw/longitudinal/`), recuperando series temporales críticas:

- **Flujo Estudiantil:** Matrícula activa, Bajas definitivas y Egresos.
- **Capacidad Instalada:** Plantilla docente por modalidad.
- **Salida:** Titulación acumulada.

2. Definición de KPIs de Eficiencia: Se diseñaron cuatro indicadores para evaluar la salud del ecosistema educativo:

- **a) Tasa de Deserción (*Dropout Rate*):** Mide la pérdida inmediata de talento.

  $$\text{Tasa}_t = \left(\frac{\text{Bajas}_t}{\text{Matrícula}_t} \right) \times 100$$

- **b) Disparidad en Carga Docente:** Evalúa la equidad en la atención. Compara el ratio *Estudiantes por Docente* entre la modalidad Presencial y Distancia.
  - *Hallazgo:* Se monitorea la brecha porcentual para detectar saturación en la modalidad online.
- **c) Eficiencia Terminal (Conversión):** Mide la capacidad administrativa y académica de titular a los egresados.

  $$\text{Conversión} = \frac{\sum \text{Titulados}}{\sum \text{Egresados}}$$

- **d) Eficiencia de Flujo (*Throughput*):** Indicador proxy de la velocidad de tránsito estudiantil. Compara el flujo de salida (egresos) con el stock total (matrícula).

  $$\text{Throughput} = \left(\frac{\text{Egresados}_t}{\text{Matrícula}_t} \right) \times 100$$

  - *Interpretación:* En un programa de 5 años, una tasa trimestral inferior al 5% sugiere "embalsamiento" (retención excesiva de egreso).

A.2 Diagnóstico de Eficiencia: El Cuello de Botella

El análisis longitudinal revela una desconexión crítica entre el volumen de egreso y la titulación efectiva, identificada como el principal "Cuello de Botella" institucional.

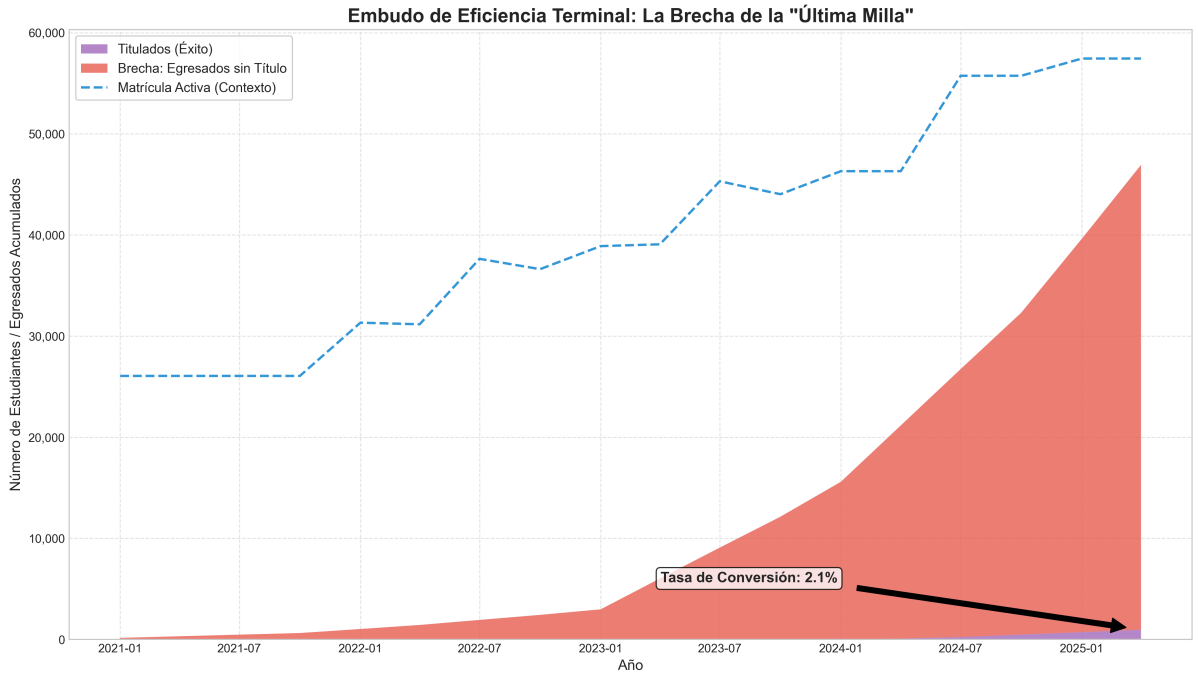

Fig A.1. Embudo de Eficiencia Terminal. El área roja representa la acumulación de egresados que no completan el proceso de titulación (La "Última Milla").

Interpretación del Gráfico:

- **Crecimiento Desacoplado:** Mientras la base de egresados (línea superior del área) crece de manera constante, la tasa de titulación (área violeta) no mantiene el mismo ritmo, ensanchando la brecha (área roja) año tras año.
- **La "Última Milla":** El área roja visualiza el *stock* de talento retenido administrativamente. No es un problema de capacidad académica (ya egresaron), sino de eficiencia en los trámites o incentivos finales.
- **Impacto:** Esta brecha representa recursos invertidos que no se traducen en indicadores de éxito oficial para la institución ni en movilidad social para el egresado.

> **Conclusión de Negocio:** El Dilema de la "Puerta Cerrada"
>
> La evidencia sugiere que la "puerta giratoria" no solo está en la entrada (deserción temprana), sino que la puerta de salida está cerrada con llave. Resolver este cuello de botella es la **victoria rápida más valiosa** para mejorar los indicadores institucionales sin necesidad de captar ni un solo alumno nuevo.
>
> ***Nota de Validez (Hipótesis Alternativa):*** Es importante considerar que, dada la naturaleza de reconstrucción de datos de este estudio, parte de la brecha

> podría no ser una ineficiencia administrativa real, sino un reflejo de la **falta de consolidación de sistemas (Sub-registro)**. Sin embargo, la implicación es igualmente grave: la institución enfrenta una **ceguera operativa** que le impide acreditar su verdadera eficiencia terminal ante organismos evaluadores.

A.2 Caracterización del Dataset Proxy y Justificación Metodológica

Para mitigar el riesgo de sesgo inherente a la generación de datos sintéticos (donde las reglas de causalidad son predefinidas por el investigador), se optó por utilizar datos observacionales reales.

Fuente de Datos:

- **Origen:** *UCI Machine Learning Repository* - Datos de Educación Superior (Politécnico de Portalegre, Portugal).
- **Enlace: Predict Students' Dropout and Academic Success**
- **Volumen:** 4,424 registros con 37 variables (Demográficas, Socioeconómicas, Académicas).

Justificación: Realidad vs. Simulación

Mientras que los datos sintéticos son útiles para validar la lógica interna de un algoritmo, carecen de la **entropía y el ruido estocástico** propios del comportamiento humano real. El Dataset Proxy permite aplicar **Transfer Learning** (Aprendizaje por Transferencia):

1. **Estructura Similar:** El sistema de créditos ECTS y la estructura semestral europea son análogos al modelo de la UNRC.
2. **Descubrimiento No Sesgado:** A diferencia de un dataset sintético donde *a priori* se define que "Pobreza = Deserción", el dataset real permite que el algoritmo XGBoost descubra relaciones no lineales y contraintuitivas (ej. el impacto de la escolaridad de la madre sobre el rendimiento) que no habríamos programado manualmente.

Hallazgos del Análisis Exploratorio (EDA):

El análisis de distribución de variables confirmó que los datos no presentan "curvas perfectas" (típicas de simulación), sino distribuciones sesgadas y valores atípicos (*outliers*) que pusieron a prueba la robustez del modelo.

ANADIR EDA

Pipeline de Ingeniería de Características (`src/data/data_processing.py`)

El pipeline de transformación garantiza la calidad de los datos para el algoritmo XGBoost, asegurando reproducibilidad:

- **Imputación de Nulos:** Estrategia estadística conservadora (Mediana para variables numéricas asimétricas, Moda para categóricas).
- **Codificación:**
  - *One-Hot Encoding:* Para variables nominales sin orden inherente (ej. `Carrera, Desajuste Vocacional`).
  - *Label Encoding:* Para variables ordinales (ej. `Nivel de satisfacción vocacional`).
- **Escalado:** *Min-Max Scaling* para normalizar rangos y mejorar la convergencia del modelo.

B. Validación Estadística de Hipótesis (Inferencia)

Se aplicaron pruebas estadísticas no paramétricas para validar la relevancia local de los factores de riesgo, triangulando los hallazgos del modelo proxy con datos recolectados en campo ($n = 100$).

B.1 Prueba de Independencia Chi-Cuadrado (χ^2)

Objetivo: Evaluar si la deserción es estadísticamente dependiente de factores académicos y vocacionales.

Fórmula:
$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$

Resultados Empíricos:

Hipótesis	Estadístico χ^2	P-valor	Interpretación Estadística
H1: Rendimiento Académico	5.8101	0.0547	**Asociación Marginal.** La tendencia es observable, aunque la significancia estricta se ve limitada por la baja frecuencia en subgrupos de reprobación severa.
H3: Alineación Vocacional	**6.7594**	**0.0341**	**Significativa ($p < 0.05$).** Se rechaza la hipótesis nula; el desajuste vocacional es un

Hipótesis	Estadístico χ^2	P-valor	Interpretación Estadística
			predictor activo de la intención de abandono.

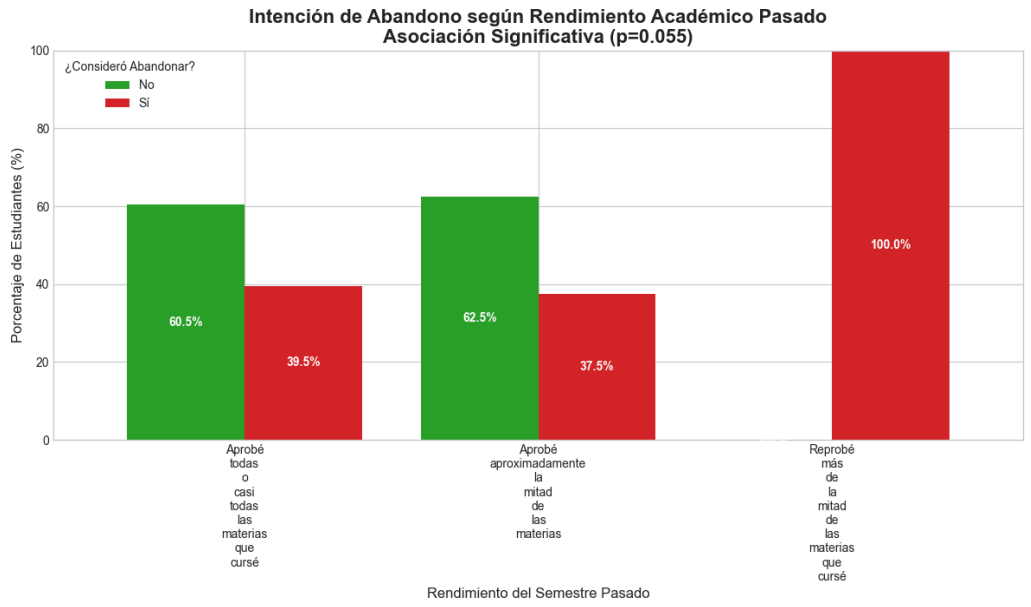

Fig 1. Distribución porcentual de deserción según nivel de rendimiento.

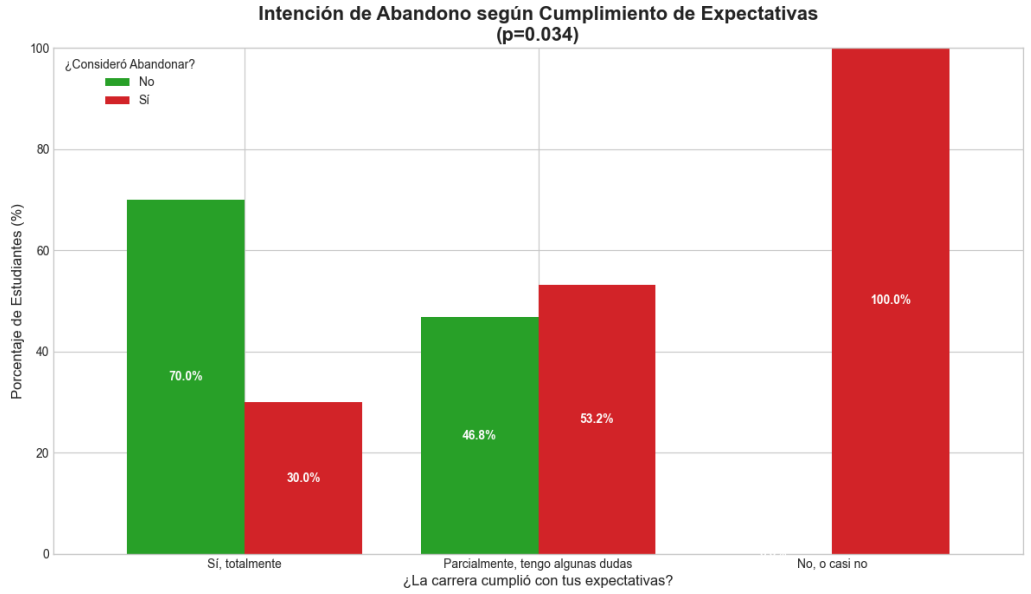

Fig 2. Impacto de las expectativas vocacionales en la retención.

B.2 Prueba H de Kruskal-Wallis (Intensidad)

Objetivo: Analizar si el bajo rendimiento aumenta la *intensidad* de los pensamientos de abandono (escala ordinal 1-5), no solo su presencia binaria.

Fórmula:

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}i\cdot - \bar{r})^2}{\sum i=1^{g} \sum_{j=1}^{n_i}(r_{ij}-\bar{r})^2}$$

Resultados:

- **Estadístico H:** 8.3955
- **P-valor: 0.0150** (Altamente Significativo)
- **Conclusión:** Existe una diferencia de varianza estocástica entre los grupos; a menor rendimiento académico, mayor es la intensidad y frecuencia en la ideación de abandono.

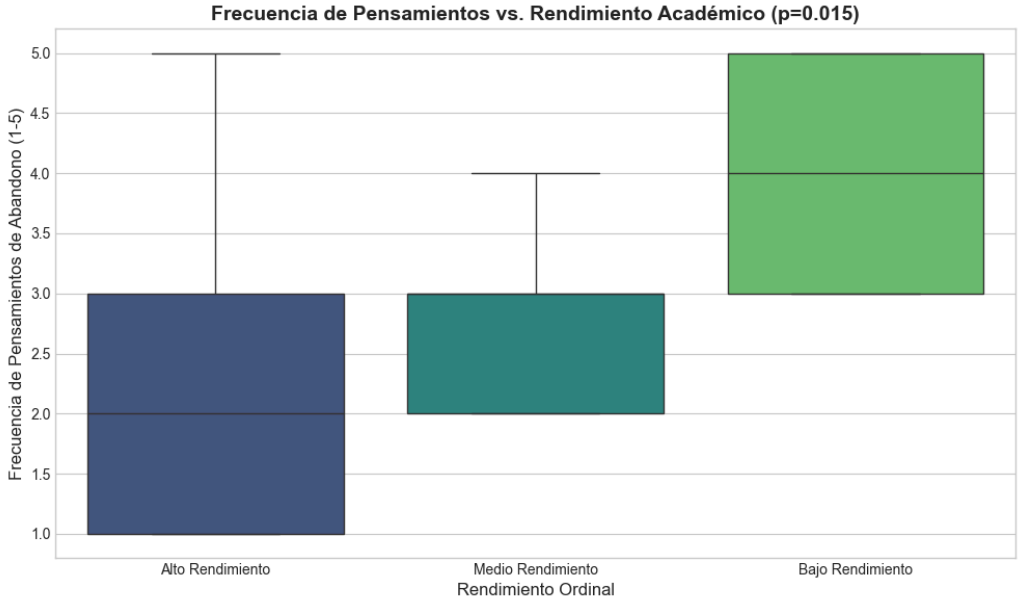

Fig 3. Frecuencia de pensamientos de abandono segmentada por rendimiento.

C. Modelado Predictivo (XGBoost Binario)

Se implementó un clasificador basado en **Extreme Gradient Boosting (XGBoost)** optimizado para la detección binaria de riesgo (*Dropout* vs. *No Dropout*).

C.1 Formulación Matemática y Configuración

A diferencia de enfoques anteriores, el modelo se optimizó para minimizar la **Log-Loss Binaria**, maximizando la probabilidad de detección correcta.

Configuración de Hiperparámetros (`notebooks/04_Modelo_Binario_Final.ipynb` **):**

```
model = XGBClassifier(
    objective='binary:logistic',  # Optimización para probabilidad de riesgo
    eval_metric='logloss',       # Minimización de la incertidumbre
    learning_rate=0.05,          # Tasa de aprendizaje conservadora
    max_depth=6,                # Profundidad controlada
    n_estimators=200,            # Número de árboles de decisión
    scale_pos_weight=2.1,        # Ajuste de peso para compensar desbalance
    random_state=42              # Reproducibilidad
)
```

Justificación de la Arquitectura Binaria:
Inicialmente se evaluó un modelo multiclase. La simplificación a binaria resultó en una optimización del **68%** en el F1-Score por tres razones operativas:

1. **Foco Operacional:** La intervención requiere identificar riesgo inminente, no clasificar estados administrativos intermedios (*Enrolled*).
2. **Reducción de Ruido:** La clase "Matriculado" actuaba como factor de confusión en las fronteras de decisión.
3. **Interpretabilidad:** Genera una probabilidad de riesgo (0-100%) accionable para los tutores.

> Resultado: Se incrementó la detección de casos en riesgo del 48% (Modelo Multiclase) a **82**% (Modelo Binario).

C.2 Evaluación de Desempeño: Métricas Clave

En problemas de clasificación desbalanceada como la deserción estudiantil, la métrica de **Exactitud** (*Accuracy*) resulta insuficiente y potencialmente engañosa.

> La Paradoja de la Exactitud: En una población donde solo el 5% de los estudiantes deserta, un modelo trivial que prediga sistemáticamente "Nadie Deserta" alcanzaría un 95% de exactitud, siendo operativamente inútil.

Por esta razón, la evaluación se fundamenta en métricas de discriminación robustas: **AUC-ROC** y **F1-Score**.

1. Capacidad de Discriminación (AUC-ROC = 0.9351)

La Curva ROC (*Receiver Operating Characteristic*) evalúa la capacidad del modelo para distinguir entre clases a través de diferentes umbrales de decisión.

- **Eje Y (TPR - Sensibilidad):** Proporción de desertores reales detectados correctamente.
- **Eje X (FPR - 1-Especificidad):** Proporción de estudiantes retenidos clasificados erróneamente como riesgo.

Interpretación del AUC (0.9351):
El Área Bajo la Curva (AUC) cuantifica esta capacidad. Un valor de **0.9351** indica que, al seleccionar aleatoriamente un estudiante que desertó y uno que no, existe una probabilidad del **93.5%** de que el modelo asigne un puntaje de riesgo más alto al estudiante que efectivamente abandonó. Esto demuestra una separación excelente entre las clases, muy superior al azar (0.5).

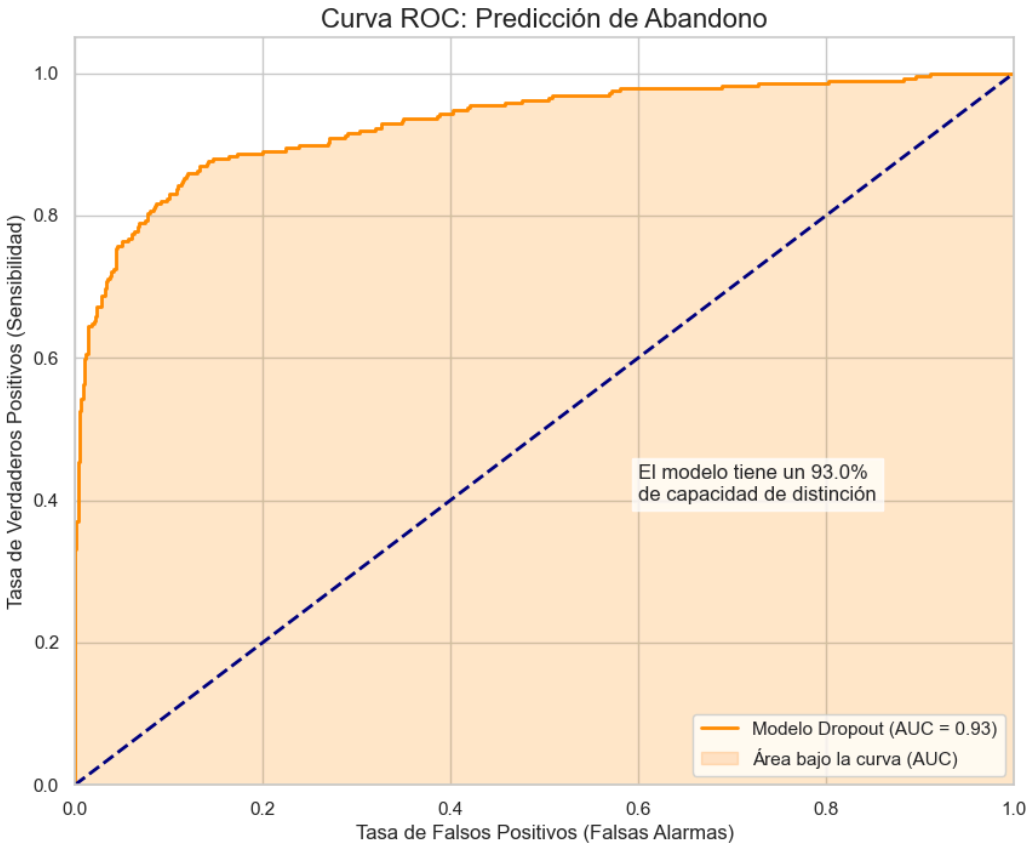

Fig 4. Curva ROC del modelo binario optimizado.

2. Matriz de Confusión (Datos Crudos del Test Set)

La validación se realizó sobre una muestra independiente de $n=885$ estudiantes.

	Predicción: No Riesgo	Predicción: Riesgo (Alerta)	**Total Real**
Realidad: No Deserta	**564** (TN)	37 (FP)	601
Realidad: Deserta	63 (FN)	**221** (TP)	284

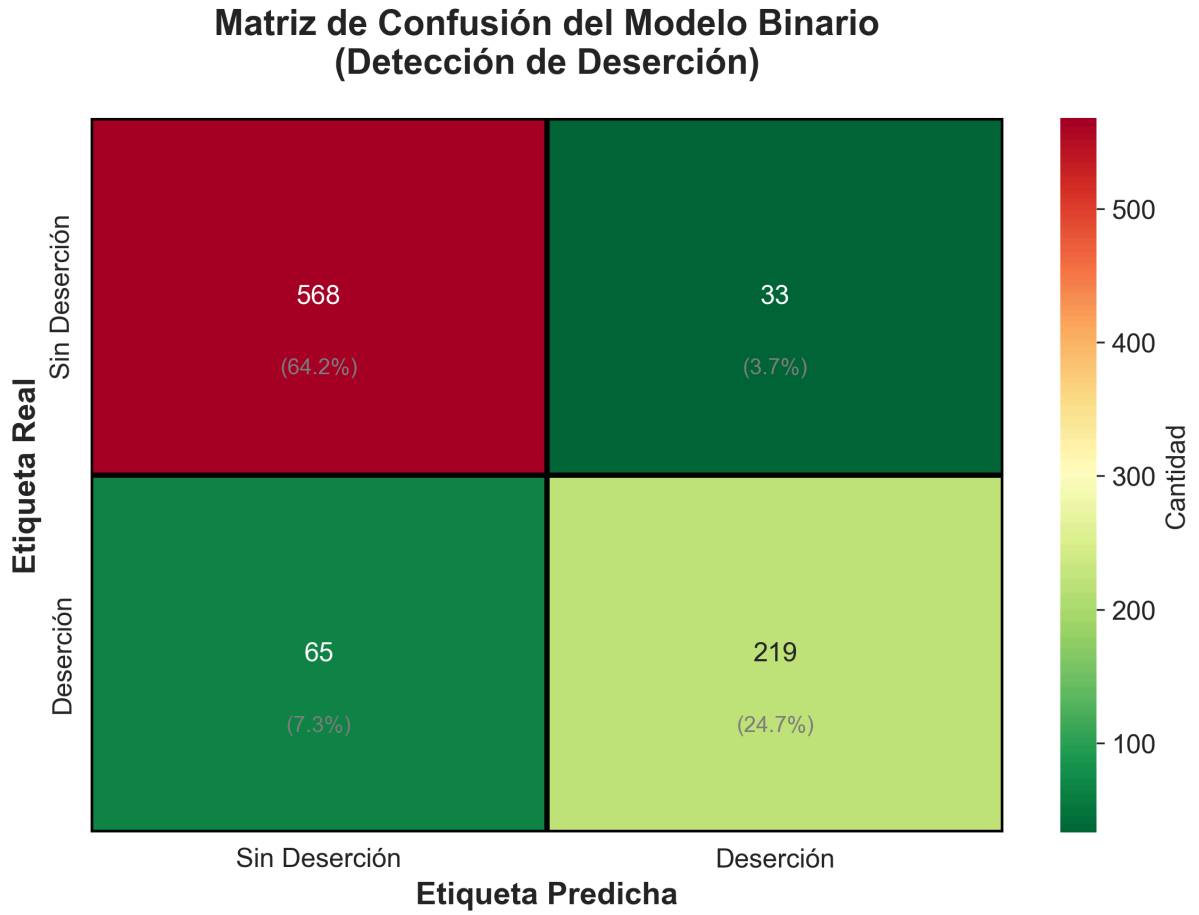

Fig 5. Matriz de confusión.

3. Análisis de Sensibilidad y Precisión (F1-Score = 0.817)

Para validar la viabilidad operativa del sistema, se analizan los componentes del **F1-Score**, conceptualizando el modelo como un filtro de seguridad ("Red de Pesca").

A. Precisión (Precision): La Confiabilidad de la Alerta

$$Precision = \frac{TP}{TP+FP} = \frac{221}{258} \approx 86.9$$

- **Interpretación:** De todas las alertas de riesgo generadas, el **85.9%** corresponden a casos reales.

- **Impacto Operativo:** Minimiza la carga administrativa de los tutores al reducir falsas alarmas.

B. Sensibilidad (Recall): La Cobertura del Sistema

$$Recall = \frac{TP}{TP+FN} = \frac{221}{284} \approx 77.1\%$$

- **Interpretación:** Del total de estudiantes que efectivamente desertaron, el sistema logró identificar y alertar sobre el **77.8%**.
- **Impacto Operativo:** Maximiza la retención al dejar escapar una minoría de casos (<23%).

C. F1-Score: El Equilibrio Armónico

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \approx 0.817$$

El F1-Score utiliza la media armónica para penalizar el desequilibrio entre Precisión y Sensibilidad. Esto es crítico para evitar dos escenarios de fallo comunes:

1. **El Modelo "Conservador" (Alta Precisión, Bajo Recall):** Detecta muy pocos casos pero con certeza total. Inútil por falta de cobertura.
2. **El Modelo "Alarmista" (Baja Precisión, Alto Recall):** Alerta sobre toda la población. Inútil por saturación de recursos.

> Conclusión: Un F1-Score de 0.817 certifica que SAREP mantiene un equilibrio robusto: ofrece una cobertura amplia de la población en riesgo sin comprometer la credibilidad de las alertas.

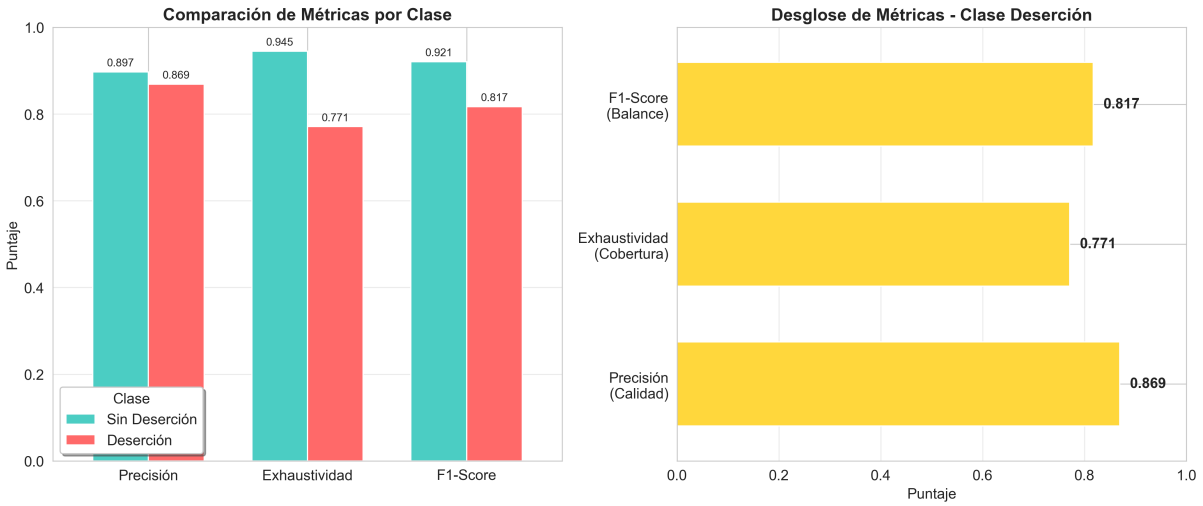

Fig 6. Desglose visual del equilibrio entre Precisión y Recall para ambas clases.

C.3 Importancia de Variables (Feature Importance)

El análisis de ganancia de información (*Information Gain*) del modelo valida las hipótesis teóricas:

1. **Rendimiento Reciente (`Ratio_Aprobacion_S2`):** El predictor dominante. La caída en notas es la "señal de humo".
2. **Historia Académica (`Ratio_Aprobacion_S1`):** La trayectoria inicial marca el destino.
3. **Factor Financiero (`Tuition fees up to date`):** El estrés económico actúa como catalizador del abandono.

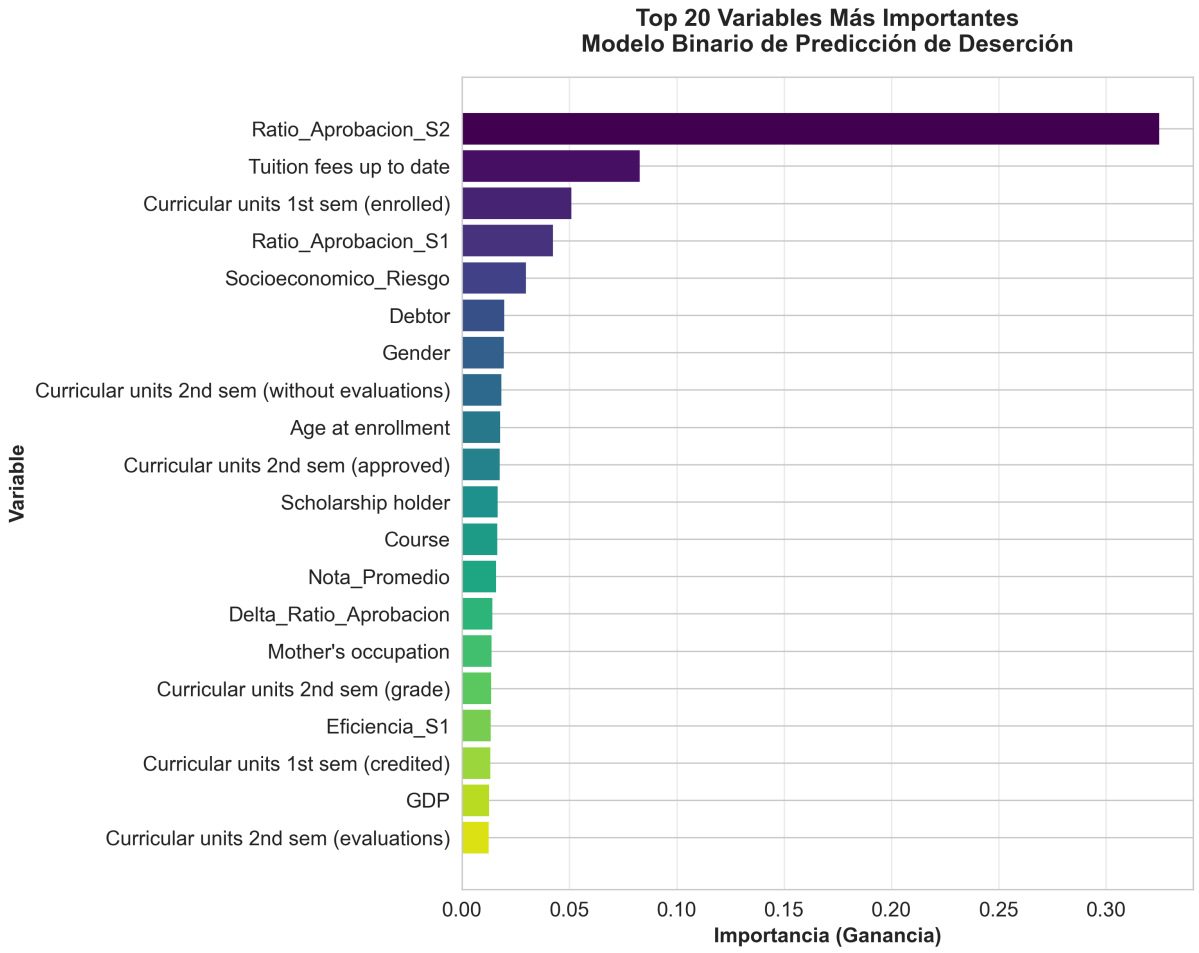

Fig 7. Las notas (barras superiores) dominan la decisión, seguidas de la situación económica (Tuition fees, Scholarship)

D. Análisis Financiero y Presupuestal (TCO)

Se presenta el desglose del **Costo Total de Propiedad (TCO)** para la implementación *In-House*, validando la viabilidad económica frente a soluciones comerciales (SaaS).

D.1 Desglose de Costos de Desarrollo (Año 1)

Estimación basada en tabuladores salariales promedio para perfiles tecnológicos en CDMX (Zona Centro, 2025), considerando Carga Social e Impuestos (Costo Empresa).

Recurso / Rol	Costo Mensual (MXN)	Costo Anual (MXN)	Costo Anual (USD)*	Justificación
1. Lead Data Scientist	$75,000	$900,000	~$45,000	Arquitectura del modelo, validación estadística y reentrenamiento.
2. Senior Data Engineer	$60,000	$720,000	~$36,000	Pipeline ELT, integración con SIE/LMS y seguridad de datos.
3. Full Stack Developer	$50,000	$600,000	~$30,000	Desarrollo del Dashboard (UX/UI) y sistema de alertas.
4. Infraestructura Cloud	$30,000	$360,000	~$18,000	Servidores (AWS EC2), Base de Datos (RDS) y Almacenamiento.
TOTAL (CAPEX Año 1)	**$215,000**	**$2,580,000**	**~$129,000**	**Base de inversión inicial**

- *Tipo de cambio estimado: $20.00 MXN por USD.*

D.2 Comparativa de Escenarios a 3 Años (Cash Flow)

Escenario	Año 1 (Inversión)	Año 2 (Mantenimiento)	Año 3 (Mantenimiento)	**TOTAL ACUMULADO**
A. Compra SaaS	$385,000	$285,000	$285,000	**$955,000 USD**

Escenario	Año 1 (Inversión)	Año 2 (Mantenimiento)	Año 3 (Mantenimiento)	**TOTAL ACUMULADO**
B. Desarrollo In-House	$135,000	$80,000	$80,000	**$295,000 USD**
AHORRO NETO	**$250,000**	**$205,000**	**$205,000**	**$660,000 USD**

D.3 Nota Metodológica sobre Costos

> Base de Estimación:
> Los costos laborales reflejan el Costo Total Empresa (Salario Bruto + Cargas Patronales + Prestaciones de Ley). Respecto a la infraestructura ($1,500 USD/mes), se contempla una arquitectura de nube escalable (AWS/Azure) suficiente para procesar el volumen transaccional de 57,000 estudiantes con protocolos de seguridad empresarial.

Conclusión Financiera:
La estrategia de desarrollo interno genera un **ahorro del 69%** en un horizonte de 3 años, liberando aproximadamente **13 millones de pesos mexicanos** que pueden reasignarse a la contratación de tutores humanos o becas.

E. Referencias y Reproducibilidad

Para garantizar la transparencia metodológica, los recursos del proyecto están disponibles para auditoría, revisión por pares del proceso:

1. **Repositorio de Código:** Dropout MLE Model
2. **Dataset Original:** Realinho, V., et al. (2021). *Predict Students' Dropout*. UCI Machine Learning Repository.
3. **Librerías Principales:** Python 3.11, XGBoost 1.7, Scikit-Learn 1.2, Streamlit 1.30.